BUILDING TYPE CLASSIFICATION FROM SOCIAL MEDIA TEXTS VIA GEO-SPATIAL TEXT MINING

Matthias Häberle[1], *Martin Werner*[2], *Xiao Xiang Zhu*[1,2]

[1] Signal Processing in Earth Observation (SiPEO), Technical University of Munich (TUM)
[2] Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR)

ABSTRACT

In this work, we present a model for building type classification from Twitter text messages (tweets) by employing geo-spatial text mining methods. First, we apply standard text pre-processing methods and convert the tweets into sentence vectors using fastText. For classification, we apply a feedforward network with two fully connected hidden layers and feed the generated sentence vectors as linguistic features. Classification results suggest that the classes are distinguishable to a certain extent with pure text even with unbalanced class distributions and a very small sample size. However, these findings also undermine, that building type classification with pure text data is a challenging task.

Index Terms— Urban Remote Sensing, Building Settlement Type, Classification, Natural Language Processing, Deep Learning, Word Embedding, Language, Social Media, Data Mining

1. INTRODUCTION

Social media platforms like Twitter are ubiquitous in our everyday lives. Twitter, for example, possesses about 336 million users worldwide [1]. As a Twitter user, one can share short utterances up to 280 characters and attach pictures, videos and connect with other content through hashtags and user mentions. Some users share their exact geo position within their message, which is called a tweet. The large user base, global coverage, and the availability of geolocation information offer a huge and rich data source which can help to discover intra-urban characteristics and structures. Also, the advent of text embedding implementations like fastText enables researches to transform text into a high dimensional feature space without losing semantic or syntactic information. Indeed, recent neural network methods support classification tasks by preserving the fine grained feature space within their many layers and produce impressive results in a broad variety of applications. In this work, we present a natural language processing based approach for building type classification using a geo referenced Twitter text messages sample of Berlin. As features for the classification task, sentence vectors have been generated out of pre-trained German word vectors. For classification, we used a feedforward neural network with two fully connected hidden layers. We applied a spatial split to separate the training and validation spatially to tackle areal overfitting.

2. RELATED WORK

Natural language processing (NLP) tries to transfer the meaning of a text written in a natural language into a data structure such that a machine can process and "interpret" the data [2]. For example, part-of-speech taggers have been proposed which work well on social media an web texts [3]. In sarcasm detection word embeddings are used to improve state-of-the-art results in this domain [4]. Certainly, machine translation is also a fundamental task in natural language processing, where researches recently achieved near human-like translation performance from Chinese to English [5]. Twitter is used in various applications fields. For example Twitter text messages have been used to produce word embeddings and to predict whether a tweet is related to the Venezuela parliamentary election in 2015 and the Philippines general election in 2016 with neural network models [6]. Geographic research areas also exploit Twitter messages. For example, annotations of OpenStreetMap[1] objects can be enriched by tweets [7]. Furthermore, the combination of remote sensing imagery and Twitter data, population density in slums in Mumbai and the quantity of social media usage have been studied [8]. In addition to the latter, it can be shown that slum dwellers not only spatially divided from the other population but also digitally. Hence, the combination of remote sensing and Twitter data delivers insightful information about urban poverty [9]. The combination of NLP methods, time series analysis, and deep learning supports urban landuse classification [10].

3. METHODS

In this section, we describe the natural language processing techniques, introduce the feedforward neural network archi-

[1] https://www.openstreetmap.org

tecture, and explain our spatial split for classification.

3.1. Text Pre-Processing

The majority of Tweets show a highly informal writing style of words and the usage of punctuation as well as other Unicode characters like emojis. To convert the tweets to a more uniform level, we set all words to lowercase, removed all punctuations, numbers and web URLs. In addition to that, we also excluded so called stopwords. Stopwords are words like *the*, *in* or *who* which usually hold no valuable information.

3.2. Sentence Vectors

To transform the pre-processed text into features, we use state of the art word embedding methods. Word embedding techniques are able to preserve semantical and syntactical features of a given corpus [2, 11, 12, 13, 14, 15]. For each word, a unique n-dimensional feature vector is computed by predicting the surrounding words of a word within a window of size w, or vice versa. In this study, we use the word embedding implementation fastText [11]. In contrast to other word embedding implementations like word2vec [12] or GloVe [13], fastText uses n-grams to create word vectors instead of full words. This approach has two advantages. First, in morphologically rich languages like German, Hebrew or Arabic [16] vector representations can be improved, and second, out of vocabulary words (OOV) and word compositions can be estimated by the n-grams [11]. We treat each tweet as a sentence. Therefore, we can compute a sentence vector of a single tweet by averaging the single word vectors belonging to the tweets words. Normally, training a word embedding from scratch consumes a high demand of computational power and huge text corpora. For this reason, we used a pre-trained German word embedding with $n = 300$ dimensions provided by the fastText development team [17]. This embedding has been trained with the entire German Wikipedia and a Common Crawl[2].

3.3. Feedforward Network and Training

For the classification task we used a feedforward network with two fully connected hidden layers (Fig. 1). The network has 300 input neurons because sentence vectors have span 300 dimensions as well. Each hidden layer has a dimension of 20 neurons and followed by a dropout layer with 0.1 magnitudes. As optimizer, we used Stochastic Gradient Descent with a learning rate of 0.1, a momentum of 0.9, a learning rate decay of $1e-6$, and activated Nesterov momentum. We applied the ReLU activation function [18] after each hidden layer and the softmax function after the output layer. The neural network has been trained for 100 epochs and a batch size of 64 samples.

[2]http://commoncrawl.org

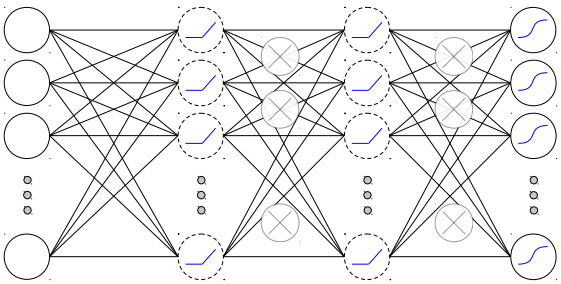

Fig. 1. Feedforward neural network

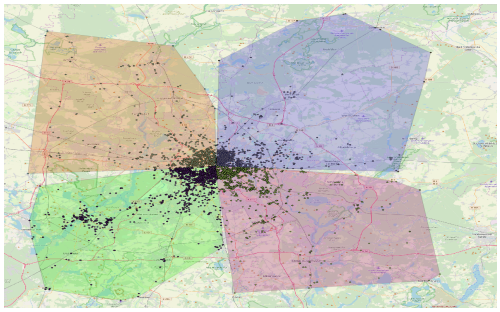

Fig. 2. Berlin spatial validation split.

3.4. Dataset and Spatial Validation Split

In order to perform a reasonable validation for our classification, we divided our Berlin sample into a north-east (NE), north-west (NW), south-east (SE) and south-west (SW) subsamples (Fig. 2). We train on three subsamples and validate on the fourth subsample. For cross validation, we rotate the subsamples that we train and validate on such that each of the splits serves as a holdout set once and average the resulting performance indicators. With this method, we want to limit spatial overfitting by showing the model a completely unseen area of Berlin.

Table 1. OpenStreetMap class distribution. The abbreviation acc. stands for "accommodation" and com. for "commercial".

	acc.	civic	com.	other	religious
NW (1)	1,766	296	665	359	22
NE (2)	4,585	228	457	303	73
SW (3)	3,358	585	1,284	566	95
SE (4)	3,398	331	589	497	61
$\sum$	13,107	1,440	2,995	1,725	251

3.4.1. OpenStreetMap Labels

We labeled our Berlin Twitter sample by assigning OpenStreetMap (OSM) building type labels to each Tweet which was done by spatial nearest neighbor join of OSM building

Table 2. Classification results. Numbers in the first column are the encoded spatial split strings (Table 1). The last number refers to the validation split. P = Precision, R = Recall, F1 = F1 score.

	accommodation			civic			commercial			other			religious		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
123-4	0.71	0.95	0.81	0.20	0.05	0.08	0.14	0.02	0.03	0.52	0.16	0.25	0.52	0.39	0.45
412-3	0.58	0.96	0.73	0.20	0.07	0.11	0.52	0.01	0.02	0.46	0.13	0.21	0.46	0.19	0.27
341-2	0.82	0.94	0.87	0.33	0.08	0.13	0.15	0.09	0.11	0.29	0.11	0.16	0.76	0.41	0.53
234-1	0.59	0.94	0.73	0.30	0.08	0.13	0.34	0.09	0.14	0.69	0.12	0.21	0.60	0.18	0.28

polygons of Berlin. Tweets, which were not within an Euclidean "distance" of 0.001 in the WGS84 coordinate space, have been removed.

3.4.2. Class Weights

We experienced unbalanced class distribution throughout our sample (Table 1) As a counter measure, we calculated class weights w by the following formulas:

$$score = \log\left(\frac{\mu * totalclasses}{classfrequency}\right) \tag{1}$$

$$w = \begin{cases} 1.0 & \text{if score} \leq \tau \\ score & \text{if score} > \tau \end{cases} \tag{2}$$

Where μ is fixed initialized with 0.15 and τ to 1.0. We calculated the class weights for all of our four training and validation splits.

4. RESULTS

As expected, the dominant class accommodation performs best (Table 2). While civic and commercial show poor results, other and religious are classified well considering that they are underrepresented.

These findings point out that the vocabulary used in other and religious could be more distinctive as in other classes and leads therefore to better classification results. For example, the German word "Kirche" (church) has been uttered eleven times in accommodation tweets, two times in other but 38 times in the religious class. In combination with other religion related words like "Gottesdienst" (service, which was written twice in accommodation class vs. eleven times in religious), tweets of the class religious could present a more unique linguistic feature space which leads to a better classification.

If one takes a look at the civic and commercial class, again, the quite poor performance is visible. It is likely that these two classes possess a more joint feature space which is not so noticeable for the classifier. Therefore, this aspect stresses the data sparsity which also states a challenge regarding the building classification task using text. While the accommodation class performs well due to its sheer dominance, other and religious make up the difference by a sufficiently unique linguistic feature space. If a class has a more subtle feature space which is not as distinct as in the religious case, additional independent data is needed to conduct building classification task.

Another issue is implicitly pointed out from the following quoted Berlin tweet:

I'm at Marienkirche in Berlin, Germany

Tweets in a metropolitan area like Berlin are rarely posted in a single language—not to mention a mixed use of languages. This means that just using pre-trained German word vectors producing the sentence vectors could "overlook" words during the process which are written in a different language then German.

5. DISCUSSION

In this study we explored if five different building classes of Berlin could be distinguished by Twitter text messages. The observed classification results indicate this. Although we found a major accommodation class, we could show that the smallest group—religious—performs quite good. This could be evidence for a different usage of linguistic patterns amongst classes which tweets are represented by sentence vectors. Thus, further investigation in that direction should be conducted. Moreover, we discovered a language distribution in our sample of 40 different languages. Therefore, dealing with more than one language within a Twitter dataset offers interesting research perspectives regarding multilanguage word embedding methods. Finally, the fusion of linguistic properties with remote sensing imagery could generate an eminent feature space to improve the building classification task further. However, it should be pointed out that building type classification on the basis of linguistic features remains a challenging task. The research with linguistic features such as word embeddings should be advanced and adapted to intra-urban language characteristics which are not necessarily monolingual or well organized.

6. ACKNOWLEDGEMENT

We gratefully acknowledge the support of the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme (grant agreement No [ERC-2016-StG-714087], Acronym: So2Sat), Helmholtz Association under the framework of the Young Investigators Group "SiPEO" (VH-NG-1018, www.sipeo.bgu.tum.de).

7. REFERENCES

[1] Twitter, "Twitter Q1 Letter to Shareholders," July 2018.

[2] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcouglu, and Pavel Kuksa, "Natural Language Processing (Almost) from Scratch," *Journal of Machine Learning Research*, vol. 12, pp. 2493–2537, 2011.

[3] Thomas Proisl, "SoMeWeTa: A Part-of-Speech Tagger for German Social Media and Web Texts," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, 2018, pp. 665–670.

[4] Aditya Joshi, Vaibhav Tripathi, Kevin Patel, Pushpak Bhattacharyya, and Mark Carman, "Are Word Embedding-based Features Useful for Sarcasm Detection?," *arXiv:1610.00883 [cs]*, Oct. 2016, arXiv: 1610.00883.

[5] Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou, "Achieving Human Parity on Automatic Chinese to English News Translation," *arXiv:1803.05567 [cs]*, Mar. 2018, arXiv: 1803.05567.

[6] Xiao Yang, Craig Macdonald, and Iadh Ounis, "Using word embeddings in Twitter election classification," *Information Retrieval Journal*, vol. 21, no. 2-3, pp. 183–207, 2018.

[7] Chen Xin, Vo Hoang, Wang Yu, and Wang Fusheng, "A framework for annotating OpenStreetMap objects using geo-tagged tweets," *Geoinformatica*, , no. 22, pp. 589–613, 2018.

[8] Martin Klotz, Michael Wurm, Xiao Xiang Zhu, and Hannes Taubenböck, "Digital deserts on the ground and from space. An experimental spatial analysis combining social network and earth observation data in megacity Mumbai," in *Joint Urban Remote Sensing Event (JURSE)*, Dubai, United Arab Emirates, 2017.

[9] Hannes Taubenböck, Jeroen Staab, Xiao Xiang Zhu, Christian Geiß, Stefan Dech, and Michael Wurm, "Are the Poor Digitally Left Behind? Indications of Urban Divides Based on Remote Sensing and Twitter Data," *ISPRS International Journal of Geo-Information*, vol. 7, no. 8, pp. 304, Aug. 2018.

[10] R. Huang, H. Taubenböck, L. Mou, and X. X. Zhu, "Classification of Settlement Types from Tweets Using LDA and LSTM," in *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, July 2018, pp. 6408–6411.

[11] Piotr Bojanowski, Edouard Grave, and Tomas Mikolov, "Enriching Word Vectors with Subword Information," *Transactions of the Association for Computational Linguistic*, vol. 5, pp. 135–146, 2017.

[12] Tomas Mikolov, Kai Chen, and Greg Corrado, "Efficient Estimation of Word Representations in Vector Space," in *Proceedings of Workshop at ICLR*, 2013.

[13] Jeffrey Pennington, Richard Socher, and Christopher D. Manning, "GloVe: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 2014, pp. 1532–1543.

[14] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin, "A Neural Probabilistic Language Model," *Journal of Machine Learning Research*, vol. 3, pp. 1137–1155, 2003.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv:1810.04805 [cs]*, Oct. 2018, arXiv: 1810.04805.

[16] Reut Tsarfaty, Djamé Seddah, Yoav Goldberg, Sandra Kübler, Marie Candito, Jennifer Foster, Yannick Versley, Ines Rehbein, and Lamia Tounsi, "Statistical Parsing of Morphologically Rich Languages (SPMRL). What, How and Wither," in *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, Los Angeles, CA, USA, 2010, pp. 1–12.

[17] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov, "Learning Word Vectors for 157 Languages," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[18] Xavier Glorot, Antoine Borders, and Yoshua Bengio, "Deep Sparse Rectifier Neural Networks," in *Proceedings of the 14 th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Fort Lauderdale, FL, USA, 2011, pp. 315–323.